Planting a SEED of Vision in Large Language Model

Year: 2023 | Citations: 121 | Authors: Yuying Ge, Yixiao Ge, Ziyun Zeng

Abstract

We present SEED, an elaborate image tokenizer that empowers Large Language Models (LLMs) with the emergent ability to SEE and Draw at the same time. Research on image tokenizers has previously reached an impasse, as frameworks employing quantized visual tokens have lost prominence due to subpar performance and convergence in multimodal comprehension (compared to BLIP-2, etc.) or generation (compared to Stable Diffusion, etc.). Despite the limitations, we remain confident in its natural capacity to unify visual and textual representations, facilitating scalable multimodal training with LLM's original recipe. In this study, we identify two crucial principles for the architecture and training of SEED that effectively ease subsequent alignment with LLMs. (1) Image tokens should be independent of 2D physical patch positions and instead be produced with a 1D causal dependency, exhibiting intrinsic interdependence that aligns with the left-to-right autoregressive prediction mechanism in LLMs. (2) Image tokens should capture high-level semantics consistent with the degree of semantic abstraction in words, and be optimized for both discriminativeness and reconstruction during the tokenizer training phase. As a result, the off-the-shelf LLM is able to perform both image-to-text and text-to-image generation by incorporating our SEED through efficient LoRA tuning. Comprehensive multimodal pretraining and instruction tuning, which may yield improved results, are reserved for future investigation. This version of SEED was trained in 5.7 days using only 64 V100 GPUs and 5M publicly available image-text pairs. Our preliminary study emphasizes the great potential of discrete visual tokens in versatile multimodal LLMs and the importance of proper image tokenizers in broader research.